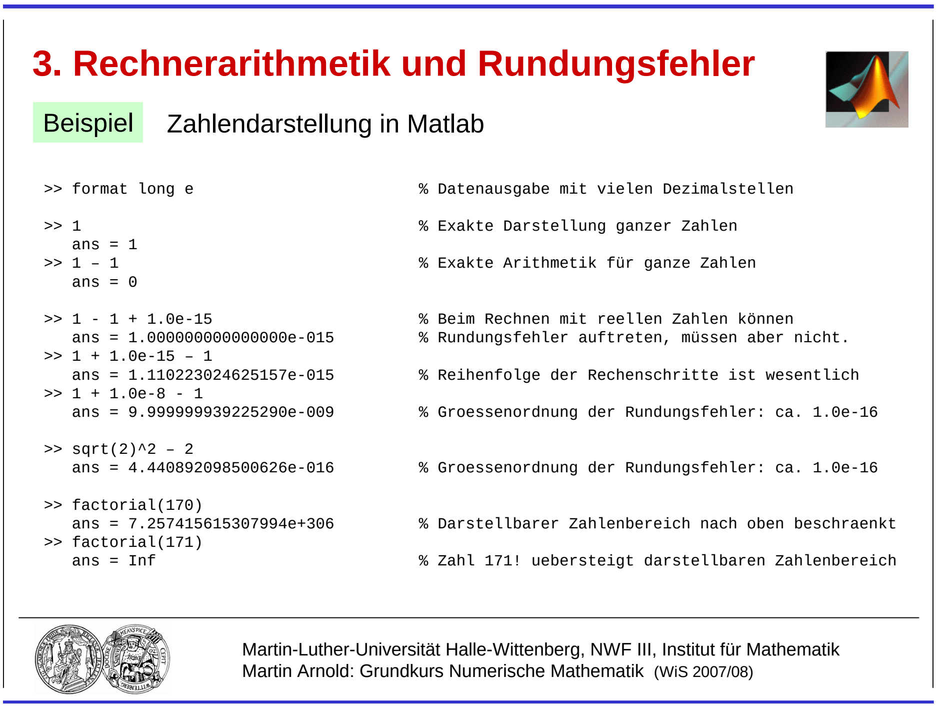


Abbildung 3.1: Rechnen in Gleitpunktarithmetik: Beispiel Matlab.

Training des Netzes Wähle $w_1, \dots, w_n$ so, dass eine große Zahl von Tests mit vorgegebenen Eingangsdaten $(x_1^{(j)}, \dots, x_n^{(j)})^\top$ und bekannten Resultaten $y^{(j)}$ möglichst gut wiedergegeben wird:

$$\sum_{i=1}^{n} x_i^{(j)} w_i \approx y^{(j)}\,, \quad (\, j = 1, \dots, m\,)$$

$\rightsquigarrow$ überbestimmtes lineares Gleichungssystem, Bestimmung von $(w_1, \dots, w_n)$ als Kleinste–Quadrate–Lösung

praktisch Lösung der Normalgleichungen oder Lösung mittels QR–Zerlegung.

3 Rechnerarithmetik und Rundungsfehler

3.1 Gleitpunktarithmetik

Bemerkung 3.1 (Gleitpunktzahlen)

a) Ganzzahlige Datentypen (INTEGER) mit exakter Arithmetik

$$-\mathrm{MaxInt} - 1\,,\ \dots\ ,\ -1,\ 0,\ 1,\ \dots\ ,\mathrm{MaxInt}\,,$$

z. B. für Indizes in Laufanweisungen.

b) *Normalisierte Gleitpunktdarstellung* (engl.: floating point numbers) zur Darstellung reeller Zahlen

$$F := \{\, y \,:\, y = \pm m * \beta^{e-t} \,\} \subset \mathbb{R}$$

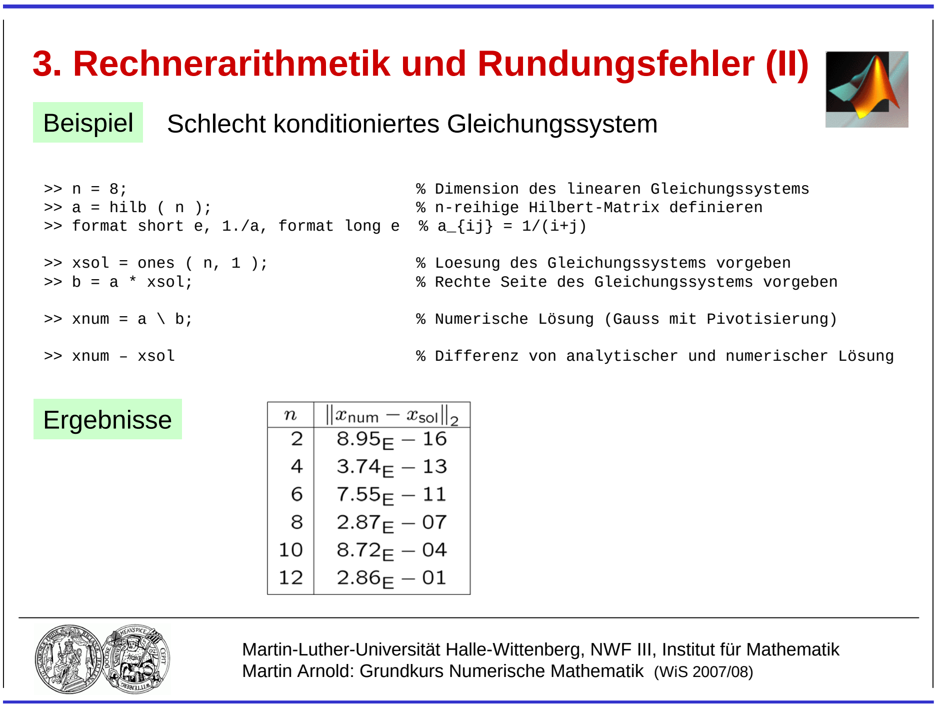


Abbildung 3.2: Rechnen in Gleitpunktarithmetik: Beispiel Hilbert–Matrix.

mit β ... Basis (meist 2, 8 oder 16),
t ... Mantissenlänge,
e ... Exponent, $e_{\min} \le e \le e_{\max}$
m ... Mantisse (ganzzahlig), $m = 0$ oder $\beta^{t-1} \le m < \beta^t$

Schreibweise $y = \pm\,\beta^e * [0.d_1 d_2 \cdots d_t]_\beta := \pm\,\beta^e \Big(\dfrac{d_1}{\beta} + \dfrac{d_2}{\beta^2} + \ldots + \dfrac{d_t}{\beta^t}\Big)$
mit Mantisse $m = d_1 d_2 \cdots d_t$ oder $m = 0$.

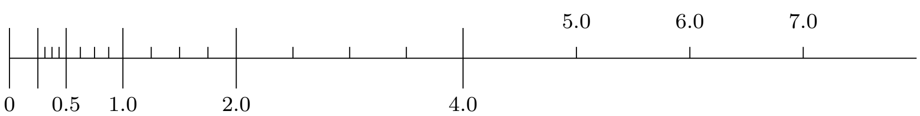
Beispiel $\beta = 2\,,\ \ t = 3\,,\ \ e_{\min} = -1\,,\ \ e_{\max} = 3$

$e=-1$ ... $e=0$ ...
$[0.100]_2$ $[0.101]_2$ $[0.110]_2$ $[0.111]_2$ $[0.100]_2$

$$F \cap [0,\infty) = \{\,0,\ 0.25,\ 0.3125,\ 0.3750,\ 0.4375,\ 0.5,\ 0.625,\ 0.750,\ 0.875,\ 1.0,\ 1.25,\ 1.50,\ 1.75,\ 2.0,\ 2.5,\ 3.0,\ 3.5,\ 4.0,\ 5.0,\ 6.0,\ 7.0\,\}$$

$[0.111]_2$
... $e=3$

Beachte: Gleitpunktzahlen sind auf der reellen Achse *nicht* gleichverteilt.



IEEE–Standard 754 [1985] Binäre Gleitpunktarithmetik, Quasi–Standard

einfache Genauigkeit (single precision)
: 4 Byte, $\beta = 2$, $t = 23$, $e_{\min} = -126$, $e_{\max} = 127$
: Zahlenbereich: $[1.2_{\mathrm{E}} - 38, 3.4_{\mathrm{E}} + 38]$

doppelte Genauigkeit (double precision)
: 8 Byte, $\beta = 2$, $t = 52$, $e_{\min} = -1022$, $e_{\max} = 1023$
: Zahlenbereich: $[2.2_{\mathrm{E}} - 308, 1.8_{\mathrm{E}} + 308]$

Abstand zweier positiver Gleitpunktzahlen x, $\tilde{x}$:

Maschinenepsilon $\texttt{eps} = \beta^{1-t}$ ... kleinste Maschinenzahl, die zu $1 = [0.10\cdots0]_\beta * \beta^1$ addiert einen von 1 verschiedenen Wert ergibt

Sind x, $\tilde{x}$ unmittelbar benachbart, so gilt

$$\frac{1}{\beta}\,\texttt{eps}\,|x| \le |x - \tilde{x}| \le \texttt{eps}\,|x|\,.$$

Bemerkung 3.2 (Rundung, Rundungsfehler)

a) Sei G die Menge aller y wie in Bemerkung 3.1, jedoch für beliebiges $e \in \mathbb{Z}$, und $\texttt{fl} : \mathbb{R} \to G$ eine Abbildung mit

$$|x - \texttt{fl}(x)| = \min_{\tilde{x}\in G} |x - \tilde{x}|\,, \quad (\,x \in \mathbb{R}\,).$$

Der Übergang $x \mapsto \texttt{fl}(x)$ heißt *runden*. $\texttt{fl}$ ist nicht eindeutig, praktisch meist: *Gerade–Zahl–Regel*, d. h., für $\tilde{x}_1$, $\tilde{x}_2 \in G$ mit $\tilde{x}_1 \ne \tilde{x}_2$ und

$$|x - \tilde{x}_1| = |x - \tilde{x}_2| = \min_{\tilde{x}\in G} |x - \tilde{x}|$$

wählt man $\texttt{fl}$ so, dass d_t geradzahlig.

b) praktisch $\texttt{fl}(x) \overset{!}{\in} F$

Exponentenüberlauf (engl.: overflow): $|\texttt{fl}(x)| > \max\{\,|y| \,:\, y \in F\,\}$

Exponentenunterlauf (engl.: underflow): $0 < |\texttt{fl}(x)| < \min\{\,|y| \,:\, y \in F\,,\ y \ne 0\,\}$

c) Zu jedem $x \in \mathbb{R}$ mit $\texttt{fl}(x) \in F$ ist

$\texttt{fl}(x) = x(1+\delta)$ mit einem δ mit $|\delta| < \varepsilon$

und

$\texttt{fl}(x) = x/(1+\bar{\delta})$ mit einem $\bar{\delta}$ mit $|\bar{\delta}| \le \varepsilon$,

wobei $\varepsilon := \frac{1}{2}\beta^{1-t}$ die *Maschinengenauigkeit* (engl.: *unit round-off*) bezeichnet:

$\varepsilon \approx 5.96_{\mathrm{E}} - 8$ (single), $\varepsilon \approx 1.11_{\mathrm{E}} - 16$ (double).

Begründung

Für $x > 0$ ist $x = \mu * \beta^{e-t}$ mit einem $\mu \in [\beta^{t-1}, \beta^t - 1]$ und $e \in [e_{\min}, e_{\max}]$. Unmittelbar benachbarte Gleitpunktzahlen: $\underline{\mu} * \beta^{e-t}$, $\bar{\mu} * \beta^{e-t}$ mit $\underline{\mu} \le \mu \le \bar{\mu}$. Es gilt

$$\begin{aligned} |x - \texttt{fl}(x)| &= \min\{\,|\mu - \underline{\mu}|, |\mu - \bar{\mu}|\,\} * \beta^{e-t} \\ &\le \frac{1}{2}\,|\bar{\mu} - \underline{\mu}| * \beta^{e-t} \le \frac{1}{2} * \beta^{e-t} = x \cdot \frac{1}{2\mu} \le x \cdot \varepsilon \end{aligned}$$

Bemerkung 3.3 (Absoluter und relativer Fehler)
a) Der *absolute Fehler* einer Größe mit Soll–Wert $\bar{\xi}$ und Ist–Wert ξ ist

$$\delta\xi := |\xi - \bar{\xi}|\,,$$

für $\bar{\xi} \neq 0$ ist der zugehörige *relative Fehler* $f_{\mathrm{rel}}(\bar{\xi}) := \dfrac{|\xi - \bar{\xi}|}{|\bar{\xi}|}$.
b) Der in Bemerkung 3.2 betrachtete Rundungsfehler erfüllt

$$f_{\mathrm{rel}}(x) = \frac{|\mathtt{fl}(x) - x|}{|x|} \leq \varepsilon\,.$$

Bemerkung 3.4 (Gleitpunktarithmetik)

Grundrechenarten $\mathtt{op} \in \{+, -, *, /\}\,, \quad \mathtt{op} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$

Problem F nicht abgeschlossen bez. $\mathtt{op}$

Anforderung („Standardmodell“)

$$\left.\begin{array}{rcl} x \mathbin{\widetilde{\mathtt{op}}} y &=& (x \mathbin{\mathtt{op}} y)(1+\delta) \\[2ex] x \mathbin{\widetilde{\mathtt{op}}} y &=& \dfrac{x \mathbin{\mathtt{op}} y}{1+\bar{\delta}} \end{array}\right\}, \quad (x, y \in F)\,,$$

mit $\delta = \delta(x, y; \mathtt{op})\,, \quad \bar{\delta} = \bar{\delta}(x, y; \mathtt{op})\,, \quad |\delta| \leq \varepsilon\,, \quad |\bar{\delta}| \leq \varepsilon\,.$

praktisch $x \mathbin{\widetilde{\mathtt{op}}} y$ „unendlich genau“ (praktisch: „mit größerer Mantissenlänge“) auswerten, anschließend runden auf nächstgelegene Gleitpunktzahl (Gerade–Zahl–Regel).

Alternativen
- Runden auf nächst kleinere bzw. nächst größere Maschinenzahl $\rightsquigarrow$ Intervallarithmetik
- „Abschneiden“ überzähliger Ziffern („chopping“)

Beispiel $\beta = 2\,, \quad t = 3\,, \quad x = \dfrac{7}{4} = [0.111]_2 * 2^1\,, \quad y = \dfrac{3}{8} = [0.110]_2 * 2^{-1}$

$$\begin{array}{rcl} x + y &=& [0.111]_2 * 2^1 + [0.110]_2 * 2^{-1} = [11.100]_2 * 2^{-1} + [0.110]_2 * 2^{-1} \\ &=& [100.010]_2 * 2^{-1} = [0.100010]_2 * 2^2 \\ x \mathbin{\tilde{+}} y &=& [0.100]_2 * 2^2 = 2.00 \end{array}$$

absoluter Fehler: $\quad |\,2.00 - \dfrac{17}{8}\,| = \dfrac{1}{8}$

relativer Fehler: $\quad \dfrac{1}{8} : \dfrac{17}{8} \approx 6\%$

Maschinengenauigkeit: $\quad \varepsilon = 2^{-3} = \dfrac{1}{8} = 12.5\%$

Bemerkung 3.5 (Rundungsfehleranalyse: Beispiel Addition)

geg.: $a, b, c \in F$
ges.: $s = a + b + c$ in Gleitpunktarithmetik

<u>$\boldsymbol{\widetilde{s} := (a \,\widetilde{+}\, b) \,\widetilde{+}\, c}$</u>

$$\begin{aligned} \widetilde{s} &= ((a \,\widetilde{+}\, b) + c)(1 + \delta_2) = ((a + b)(1 + \delta_1) + c)(1 + \delta_2) \\ &= s + (a + b)\delta_1 + (a + b + c)\delta_2 + (a + b)\delta_1\delta_2 \mathrel{\underset{\cdot}{=}} s + (a + b)\delta_1 + s \cdot \delta_2 \end{aligned}$$

mit $|\delta_1|, |\delta_2| \le \varepsilon$. Terme höherer Ordnung werden vernachlässigt („ $\underset{\cdot}{=}$ ”).

$$f_{\text{rel}}(s) \mathrel{\underset{\cdot}{=}} \left| \delta_2 + \frac{a + b}{a + b + c}\, \delta_1 \right| \le \Big(1 + \left| \frac{a + b}{a + b + c} \right| \Big)\, \varepsilon$$

Beachte Fehler in Zwischenergebnissen (δ_1) können verstärkt werden, ebenso auch Fehler in Ausgangsdaten.

kritisch $|a + b + c| \ll |a + b|$

<u>$\boldsymbol{\widetilde{\widetilde{s}} := a \,\widetilde{+}\, (b \,\widetilde{+}\, c)}$</u>

$$\widetilde{\widetilde{s}} \mathrel{\underset{\cdot}{=}} s + (b + c)\delta_3 + s \cdot \delta_4 \quad \text{mit} \quad |\delta_3|, |\delta_4| \le \varepsilon\,,$$

$$f_{\text{rel}}(s) \mathrel{\underset{\cdot}{=}} \left| \delta_4 + \frac{b + c}{a + b + c}\, \delta_3 \right| \le \Big(1 + \left| \frac{b + c}{a + b + c} \right| \Big)\, \varepsilon$$

Beachte Gleitpunktoperationen sind in der Regel weder assoziativ noch kommutativ.

Beispiel 3.6 (Addition in Gleitpunktarithmetik)

geg.: $\beta = 2\,, \;\; t = 3$

$$a = [0.111]_2 * 2^0 = \frac{7}{8}\,, \quad b = -[0.110]_2 * 2^0 = -\frac{6}{8}\,, \quad c = [0.110]_2 * 2^{-2} = \frac{3}{16} \in F$$

$$\begin{aligned} \widetilde{s} &= (\,[0.111]_2 * 2^0 \;\widetilde{-}\; [0.110]_2 * 2^0\,) \;\widetilde{+}\; [0.110]_2 * 2^{-2} \\ &= [0.001]_2 * 2^0 \;\widetilde{+}\; [0.110]_2 * 2^{-2} = [0.100]_2 * 2^{-2} \;\widetilde{+}\; [0.110]_2 * 2^{-2} \\ &= [1.01]_2 * 2^{-2} = [0.101]_2 * 2^{-1} = \frac{5}{16} = s\,, \;\; \text{exaktes Ergebnis} \end{aligned}$$

$$\begin{aligned} \widetilde{\widetilde{s}} &= [0.111]_2 * 2^0 \;\widetilde{+}\; (\,-[0.110]_2 * 2^0 \;\widetilde{+}\; [0.110]_2 * 2^{-2}\,) \\ &= [0.111]_2 * 2^0 \;\widetilde{-}\; [0.100]_2 * 2^0 = [0.011]_2 * 2^0 = [0.110]_2 * 2^{-1} = \frac{3}{8} \end{aligned}$$

Ergebnis: $|\widetilde{\widetilde{s}} - s| = \dfrac{1}{16}$, relativer Fehler 20%

Relativer Fehler in $b \widetilde{+} c$: $\dfrac{1}{8} = 12.5\%$

Verstärkung im Endergebnis $\widetilde{\widetilde{s}}$ wegen $\left|\dfrac{b+c}{a+b+c}\right| = \dfrac{9}{5} = 1.8$

Bemerkung 3.7 (Auslöschung)

Problem Subtraktion annähernd gleich großer Zahlen in Gleitpunktarithmetik
geg.: $a, b \in \mathbb{R}$, ges.: $a - b$

$$\widetilde{a} := \mathtt{fl}(a) = a\,(1 + \delta_a)\,,\ \widetilde{b} := \mathtt{fl}(b) = b\,(1 + \delta_b) \quad \text{mit } |\delta_a|, |\delta_b| \leq \varepsilon$$

$$\begin{aligned} \mathtt{fl}(a-b) &= \widetilde{a} \,\widetilde{-}\, \widetilde{b} = (\widetilde{a} - \widetilde{b})\,(1 + \delta_-) \quad \text{mit } |\delta_-| \leq \varepsilon \\ &\doteq a - b + \delta_- \cdot (a-b) + \delta_a \cdot a - \delta_b \cdot b \end{aligned}$$

$$f_{\text{rel}}(a-b) \doteq \left| \frac{a}{a-b}\,\delta_a - \frac{b}{a-b}\,\delta_b + \delta_- \right| \leq \left(1 + \frac{|a| + |b|}{|a-b|}\right) \varepsilon$$

Relative Fehler δ_a, δ_b in den Ausgangsdaten können drastisch verstärkt werden, falls $|a - b| \ll |a|, |b|$, insbesondere für $a \approx b$.

Beispiel $\beta = 2$, $a = \dfrac{3}{5}$, $b = \dfrac{4}{7}$

$\boldsymbol{t = 5}$ $\widetilde{a} = [0.10011]_2 * 2^0$, $\widetilde{b} = [0.10010]_2 * 2^0$

$\delta_a \approx 0.010$, $\delta_b \approx 0.016$, $\varepsilon \approx 0.031$

$\widetilde{a} \,\widetilde{-}\, \widetilde{b} = [0.10000]_2 * 2^{-4} = \dfrac{1}{32}$

absoluter Fehler: $\dfrac{1}{32} - \dfrac{1}{35}$, relativer Fehler: 8.6%

$\boldsymbol{t = 3}$ $\widetilde{a} = \widetilde{b} = [0.100]_2 * 2^0$

$\delta_a \approx 0.042$, $\delta_b \approx 0.094$, $\varepsilon = 0.125$

$\widetilde{a} \,\widetilde{-}\, \widetilde{b} = 0$, relativer Fehler: 100%

Führende Ziffern $[0.1001\cdots]_2$ in a und b sind gleich und werden bei Subtraktion „ausgelöscht“ $\Rightarrow$ „Auslöschung“.

Faustregel Vermeide – falls möglich – die Subtraktion annähernd gleich großer Zahlen in numerischen Algorithmen.

Strategien und Tricks

a) Unvermeidbare Subtraktionen annähernd gleich großer Zahlen möglichst an den Anfang des Algorithmus stellen.

b) Konjugierte Wurzelausdrücke

- $$\begin{aligned} \sqrt{1+x}-\sqrt{1-x} &= \frac{(\sqrt{1+x}-\sqrt{1-x})(\sqrt{1+x}+\sqrt{1-x})}{\sqrt{1+x}+\sqrt{1-x}} \\ &= \frac{2x}{\sqrt{1+x}+\sqrt{1-x}}\,, \qquad |x| \ll 1 \end{aligned}$$

- $x^2+px+q=0 \;\Rightarrow\; x_{1,2} = -\frac{p}{2} \pm \sqrt{\frac{p^2}{4}-q}$

  Auslöschung für $|q| \ll 1$, deshalb für $p \neq 0$

$$x_1 := -\frac{p}{2} - \operatorname{sgn}(p)\sqrt{\frac{p^2}{4}-q} \quad \text{mit} \quad \operatorname{sgn}(p) := \begin{cases} 1 & \text{für } p>0\,, \\ 0 & \text{für } p=0\,, \\ -1 & \text{für } p<0\,, \end{cases}$$

$$x_2 := \frac{q}{x_1} \qquad \text{(VIETAscher Wurzelsatz)}$$

c) Analytische Umformungen, z. B. Reihenentwicklungen (vgl. Abb. 3.3)

$$\frac{1-\cos x}{x} = \frac{1}{x}\Big(1-(1-\frac{x^2}{2}+\frac{x^4}{24}\mp\ \dots)\Big) = \frac{x}{2}(1-\frac{x^2}{12}\pm\ \dots)$$

Fehler der Approximation $\frac{1-\cos x}{x} \approx \frac{x}{2}$ betragsmäßig beschränkt durch $\frac{|x|}{2}\cdot\frac{x^2}{12}$
(Reihenrest alternierender Reihen, Satz von Leibniz)

3.2 Vektor- und Matrixnormen

Definition 3.8 (Vektornorm)

Eine Abbildung $\|.\| : \mathbb{R}^n \to \mathbb{R}$ *heißt* Vektornorm auf $\mathbb{R}^n$, *falls*

1. $\|x\| \geq 0\,,\ (x \in \mathbb{R}^n)$ *und* $\big(\,\|x\|=0 \;\Leftrightarrow\; x=0\,\big)$ *(Positivität),*
2. $\|\alpha x\| = |\alpha|\,\|x\|\,,\ (\alpha \in \mathbb{R},\ x \in \mathbb{R}^n)$ *(Homogenität),*
3. $\|x+y\| \leq \|x\|+\|y\|\,,\ (x,y \in \mathbb{R}^n)$ *(Dreiecksungleichung).*

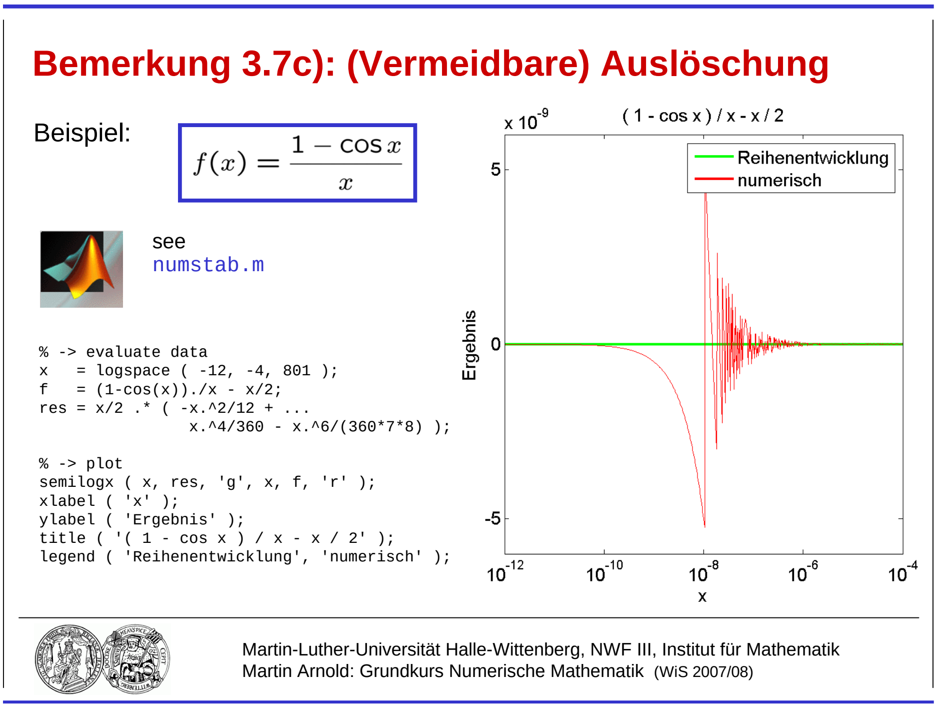


Abbildung 3.3: Analytische Umformungen zur Vermeidung von Auslöschung.

Beispiel 3.9 (Vektornorm)

a) $\|x\|_2 := \sqrt{\sum_{i=1}^{n} x_i^2}$ ... Euklidische Vektornorm

$\|x\|_1 := \sum_{i=1}^{n} |x_i|$ ... 1–Norm

$\|x\|_\infty := \max_{i=1,\dots,n} |x_i|$ ... Maximumnorm, ∞–Norm

b) Kugeln im $\mathbb{R}^n$: $\{\, x \,:\, \|x\| \le 1 \,\}$, vgl. Abb. 3.4.

Bemerkung 3.10 (Eigenschaften von Vektornormen)

a) Jedes Skalarprodukt $\langle\, .\,,. \,\rangle$ in $\mathbb{R}^n$ erzeugt eine Vektornorm in $\mathbb{R}^n$:

$$\|x\| := \sqrt{\langle\, x, x \,\rangle}$$

mit $|\langle\, x, y \,\rangle| \le \|x\| \cdot \|y\|\,,\ (\, x, y \in \mathbb{R}^n\,)$ ... Cauchy–Schwarzsche Ungleichung.

b) Auf $\mathbb{R}^n$ sind sämtliche Vektornormen *äquivalent*, d. h., zu beliebig vorgegebenen Vektornormen $\|.\|_p$, $\|.\|_q$ gibt es Konstanten $\underline{c}$, $\overline{c} > 0$ mit

$$\underline{c}\|x\|_q \le \|x\|_p \le \overline{c}\|x\|_q\,,\ \ (\, x \in \mathbb{R}^n\,)\,.$$

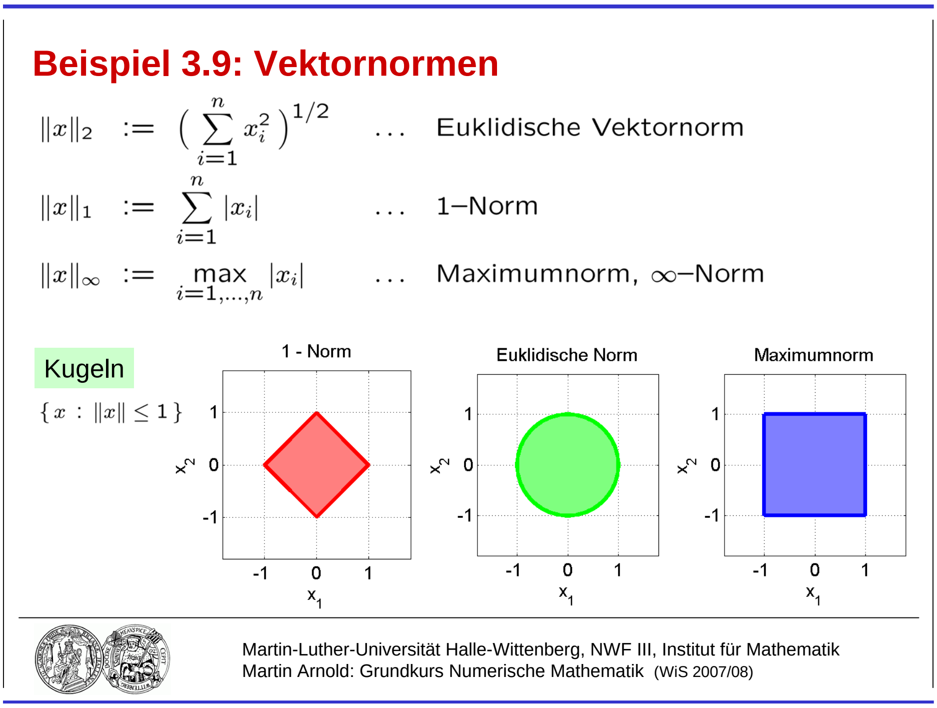


Abbildung 3.4: Einheitskugeln im $\mathbb{R}^2$.

Definition 3.11 (Matrixnorm)

a) Eine Abbildung $\|.\| : \mathbb{R}^{m\times n} \to \mathbb{R}$ *heißt* Matrixnorm, *falls*

1. $\|A\| \ge 0\,,\ (A \in \mathbb{R}^{m\times n})$ *und* $\big(\, \|A\| = 0 \;\Leftrightarrow\; A = 0 \,\big)$ *(Positivität),*
2. $\|\alpha A\| = |\alpha| \cdot \|A\|\,,\ (\alpha \in \mathbb{R},\ A \in \mathbb{R}^{m\times n})$ *(Homogenität),*
3. $\|A + B\| \le \|A\| + \|B\|\,,\ (A, B \in \mathbb{R}^{m\times n})$ *(Dreiecksungleichung).*

b) Eine Matrixnorm $\|.\|$ *heißt* submultiplikativ, *falls*

$$\|AB\| \le \|A\| \cdot \|B\|\,,\quad (A \in \mathbb{R}^{m\times n}\,,\ B \in \mathbb{R}^{n\times p})\,.$$

c) Eine submultiplikative Matrixnorm $\|.\|$ *heißt* verträglich *(auch:* konsistent*) mit einer vorgegebenen Vektornorm* $\|.\|$*, falls*

$$\|Ax\| \le \|A\| \cdot \|x\|\,,\quad (A \in \mathbb{R}^{m\times n}\,,\ x \in \mathbb{R}^{n})\,.$$

Beispiel 3.12 (Frobeniusnorm)

$$\|A\|_{\mathrm{F}} := \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}^2} \qquad \dots \text{ Frobeniusnorm}$$

- Submultiplikative Matrixnorm, verträglich mit $\|.\|_2$:

$$\|Ax\|_2^2 = \sum_{i=1}^{m} \Big(\sum_{j=1}^{n} a_{ij} x_j \Big)^2 \le \sum_{i=1}^{m} \Big(\sum_{j=1}^{n} a_{ij}^2 \Big) \Big(\sum_{j=1}^{n} x_j^2 \Big) = \|A\|_{\mathrm{F}}^2 \cdot \|x\|_2^2$$

- $\|I_n\| = \sqrt{n}$

Satz 3.13 (Zugeordnete Matrixnorm)
Zu einer vorgegebenen Vektornorm $\|.\|$ *wird durch*

$$A \mapsto \|A\| := \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|} = \sup_{\|x\|=1} \|Ax\|$$

eine submultiplikative, mit $\|.\|$ *verträgliche Matrixnorm definiert, die* der Vektornorm $\|.\|$ zugeordnete Matrixnorm. *Es gilt* $\|I_n\| = 1$ *.*

Beweis vgl. Huckle/Schneider, Anhang B.2, z. B.

$$\|I_n\| = \sup_{x \neq 0} \frac{\|I_n x\|}{\|x\|} = 1\,. \quad \blacksquare$$

Beispiel 3.14 (Zugeordnete Matrixnorm)
a) *Zeilensummennorm*

$$\|A\|_\infty := \max_{i=1,\ldots,m} \sum_{j=1}^{n} |a_{ij}|$$

ist $\|x\|_\infty$ zugeordnet, denn

- $\|Ax\|_\infty = \max\limits_{i=1,\ldots,m} \left| \sum\limits_{j=1}^{n} a_{ij} x_j \right| \leq \max\limits_{i=1,\ldots,m} \sum\limits_{j=1}^{n} |a_{ij}|\,|x_j| \leq \|A\|_\infty \cdot \|x\|_\infty$,
- zu einem $i_0 \in \{1, \ldots, m\}$ mit $\sum\limits_{j=1}^{n} |a_{i_0,j}| = \max\limits_{i=1,\ldots,m} \sum\limits_{j=1}^{n} |a_{ij}|$ wählt man $x \in \mathbb{R}^n$ so, dass $\|x\|_\infty = 1$ und $x_j = 1$, falls $a_{i_0,j} > 0$, $x_j = -1$, falls $a_{i_0,j} < 0$

$$\Rightarrow \quad \|Ax\|_\infty \geq \sum_{j=1}^{n} |a_{i_0,j} x_j| = \sum_{j=1}^{n} |a_{i_0,j}| = \|A\|_\infty\,.$$

b) *Spaltensummennorm*

$$\|A\|_1 := \max_{j=1,\ldots,n} \sum_{i=1}^{m} |a_{ij}| \quad \text{ist } \|x\|_1 \text{ zugeordnet.}$$

c) *Spektralnorm*

$$\|A\|_2 := \max_{i=1,\ldots,n} \sqrt{\lambda_i(A^\top A)}$$

Für orthogonale Matrizen $U \in \mathbb{R}^{n \times n}$, also $U^\top U = I_n$, ist $\lambda_i(U^\top U) = 1$ und $\|U\|_2 = 1$.

3.3 Kondition und Stabilität

Bemerkung 3.15 (Eingabefehler und Fehler im Ergebnis)

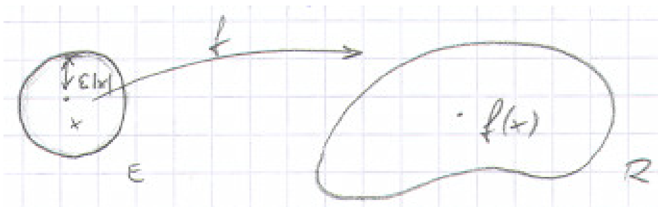
analytisch Eingabe x $\rightarrow$ Algorithmus / Berechnungsvorschrift
$\rightarrow$ Resultat $f(x)$

numerisch Fehler im Resultat entstehen durch
– Eingabefehler
– Fehler im Algorithmus

Numerische Eingabe $x \in F$ repräsentiert

Eingabemenge $E = \{\, \tilde{x} \in \mathbb{R} \,:\, \mathtt{fl}(\tilde{x}) = x \,\}$

Resultatmenge $R = f(E) := \{\, f(\tilde{x}) \,:\, \tilde{x} \in E \,\}$



„Kondition“ Maß für das Verhältnis von R zu E

Ziel Fehler in der Berechnungsvorschrift sollen Menge R nicht deutlich vergrößern

Fehler im Ergebnis $y = f(x)$:

$$\delta_y := f(x + \delta_x) - f(x) \doteq f'(x)\, \delta_x$$

Relativer Fehler:

$$\frac{\|\delta_y\|}{\|y\|} \doteq \frac{\|f'(x)\, \delta_x\|}{\|y\|} \leq \frac{\|x\|\, \|f'(x)\|}{\|y\|} \cdot \frac{\|\delta_x\|}{\|x\|}$$

Definition 3.16 (Konditionszahl)
Zu einem Problem $x \mapsto f(x)$ *heißt*

$$\mathrm{cond}_x := \frac{\|x\|\, \|f'(x)\|}{\|f(x)\|}$$

Konditionszahl. *Das Problem ist* gut konditioniert, *wenn* cond_x *klein ist, und* schlecht konditioniert *für große Konditionszahlen* cond_x.

Beispiel 3.17 (Kondition)

Exponentialfunktion $x \mapsto \mathrm{e}^x,\ \mathrm{cond}_x = \left|\dfrac{x\, \mathrm{e}^x}{\mathrm{e}^x}\right| = |x|$, gut konditioniert für $|x| \lessapprox 1$.

Logarithmus $x \mapsto \ln x,\ \mathrm{cond}_x = \left|\dfrac{x \cdot \frac{1}{x}}{\ln x}\right| = \dfrac{1}{|\ln x|}$, sehr schlecht konditioniert für $x \approx 1$.

Gute / schlechte Kondition

Polynomnullstellen häufig schlecht konditioniertes Problem

Beispiel: $\pi(t) = t^4 - 8t^3 + 24t^2 - 32t + 15.999\,999\,99 = (t-2)^4 - 10^{-8}$

$t_{1,2} = 2 \pm 0.01\,, \quad t_{3,4} = 2 \pm 0.01i$

Relativer Fehler bei Darstellung von 15.999 999 99 durch 16.0: $\varepsilon_x = 6.25 \cdot 10^{-10}$, z. B. bei Maschinenengenauigkeit $\varepsilon = 8 \cdot 10^{-10} \quad \Rightarrow \quad t_{1,2,3,4} = 2.0$

$$\mathrm{cond}_{t_{1,2}} = \frac{0.01/2.01}{6.25 \cdot 10^{-10}} \approx 8.0 \cdot 10^6$$

Bemerkung 3.18 (Berechnungsvorschrift)
Zum mathematischen Problem $x \mapsto f(x)$ sei die Abbildung $x \mapsto \tilde{f}(x)$ gegeben zur Berechnung von $f(x)$ in Gleitpunktarithmetik (u. a. auch Reihenfolge der Rechenoperationen festgelegt) $\rightsquigarrow$ *Berechnungsvorschrift*

Beispiel $f(x) = 1 - \sqrt{1 - x^2}$, für $|x| \ll 1$ gut konditioniert, $\mathrm{cond}_x \approx 2$.

Berechnungsvorschrift 1: $\tilde{f}(x) := 1 - \left(\sqrt{1 - (x^2)}\right)$

Berechnungsvorschrift 2: $\tilde{f}(x) := \dfrac{(x^2)}{\left(1 + \left(\sqrt{1 - (x^2)}\,\right)\right)}$

Definition 3.19 (Numerische Stabilität)
Zu einem gut konditionierten Problem $x \mapsto f(x)$ heißt eine Berechnungsvorschrift $x \mapsto \tilde{f}(x)$ numerisch stabil, *wenn die relativen Eingabefehler durch die Berechnungsvorschrift nicht vergrößert werden, und* numerisch instabil *sonst.*

Bemerkung 3.20 (Lineare Gleichungssysteme: Kondition und Stabilität)
a) Betrachte zu gegebener regulärer Matrix $A \in \mathbb{R}^{n \times n}$ und gegebenem $b \in \mathbb{R}^n$ die Lösung

des linearen Gleichungssystems $Ax = b$ als Abbildung $b \mapsto x := A^{-1}b$ mit gestörten Eingangsdaten b und exakten Matrizen A, A^{-1} $\Rightarrow$

$$\begin{aligned} &A(x + \delta_x) = b + \delta_b\,, \quad Ax = b \\ &\|\delta_x\| = \|A^{-1}\delta_b\| \le \|A^{-1}\|\,\|\delta_b\| \\ &\|b\| = \|Ax\| \le \|A\|\,\|x\| \end{aligned}$$

Ergebnis $\quad \dfrac{\|\delta_x\|}{\|x\|} \le \|A\| \cdot \|A^{-1}\| \cdot \dfrac{\|\delta_b\|}{\|b\|}$

- Empfindlichkeit der Lösung gegenüber Störungen in den Eingangsdaten wird beschrieben durch die *Konditionszahl* $\operatorname{cond}(A) := \|A\| \cdot \|A^{-1}\|$.
- Ergebnis lässt sich übertragen auf Empfindlichkeit gegenüber Störungen in A.
- Wegen $1 \le \|I_n\| = \|A \cdot A^{-1}\| \le \|A\| \cdot \|A^{-1}\|$ gilt stets $\operatorname{cond}(A) \ge 1$.

  Gut konditioniert: $\quad \operatorname{cond}(A) \lessapprox 10^3$

  Schlecht konditioniert: $\quad \operatorname{cond}(A) \gg 10^6$

Beispiel 1

$$A = \begin{pmatrix} 1 & 0 \\ 0 & \varepsilon \end{pmatrix}, \quad A^{-1} = \begin{pmatrix} 1 & 0 \\ 0 & 1/\varepsilon \end{pmatrix} \quad \text{mit } 0 < \varepsilon \ll 1$$

$\Rightarrow \quad \operatorname{cond}_2(A) = 1 \cdot \dfrac{1}{\varepsilon} = \dfrac{1}{\varepsilon} \to \infty$ für $\varepsilon \to 0$, Fehlerverstärkung um Faktor $1/\varepsilon$ möglich

Beispiel 2

Hilbert–Matrizen $H^{(n)} = (h_{ij}^{(n)})_{i,j} \in \mathbb{R}^{n\times n}$ mit $h_{ij}^{(n)} := 1/(i+j-1)$, $(i,j = 1,\ldots,n)$.

Zu $b^{(n)} \in \mathbb{R}^n$, $b^{(n)} = (b_i^{(n)})_i$ mit $b_i^{(n)} := \displaystyle\sum_{j=1}^{n} \frac{1}{i+j-1}$ ist die Lösung $x^{(n)}$ des linearen Gleichungssystems $H^{(n)}x^{(n)} = b^{(n)}$ gegeben durch $x^{(n)} = (1, 1, \ldots, 1)^\top$.

Fehler der mit Matlab ($\varepsilon = 1.1_{\mathrm{E}} - 16$) berechneten Lösung $\tilde{x}^{(n)}$:

n	$\|\tilde{x}^{(n)} - x^{(n)}\|_2$	$\operatorname{cond}_2(H^{(n)})$
2	9.0e-16	1.9e+01
4	4.6e-13	1.6e+04
6	3.5e-10	1.5e+07
8	1.3e-08	1.5e+10
10	3.0e-04	1.6e+13
12	1.6e+00	1.7e+16

b) Für orthogonale Matrizen $Q \in \mathbb{R}^{n\times n}$ ist $Q^{-1} = Q^\top$ und $\|Q\|_2 = \|Q^\top\|_2 = 1$
$\Rightarrow \quad \operatorname{cond}_2(Q) = 1$. Operationen mit orthogonalen Matrizen lassen die Kondition einer Matrix unverändert: $A = QR \;\Rightarrow\; \operatorname{cond}_2(R) = \operatorname{cond}_2(A)$.

c) Realisierung des Gauß–Algorithmus in Gleitpunktarithmetik:

Fehlerschranke hängt linear ab von $\max\limits_{i,k} |l_{ik}|$.

Spaltenpivotisierung: $|l_{ik}| \leq 1 \rightsquigarrow$ kleine Fehlerschranke

Numerische Stabilität: numerische Lösung $\tilde{x}$ erfüllt

$$(A + \delta_A)\tilde{x} = b$$

mit

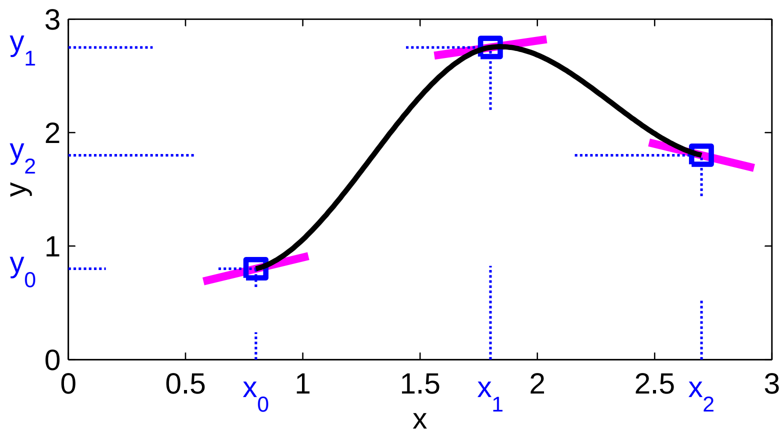
$$\frac{\|\delta_A\|_\infty}{\|A\|_\infty} \mathrel{\dot{\leq}} 8n^3 \cdot \frac{\max_{i,j,k} |a_{ij}^{(k)}|}{\max_{i,j} |a_{ij}|} \varepsilon \,.$$

4 Interpolation (II)

Bemerkung 4.1 (Stückweise Hermite–Interpolation)

geg.: $r+1$ Stützstellen $x_0, x_1, \ldots, x_r$

Stützwerte (y_k, y_k'), $(k = 0, 1, \ldots, r)$



Definiert man die interpolierende Funktion Φ stückweise durch Hermite–Interpolationspolynome $\Phi\big|_{[x_{i-1},x_i]}$, $(i = 1, \ldots, r)$ mit Interpolationsbedingungen

$$\Phi(x_{i-1}) = y_{i-1}\,, \quad \Phi'(x_{i-1}) = y_{i-1}'\,, \quad \Phi(x_i) = y_i\,, \quad \Phi'(x_i) = y_i'\,, \quad (i = 1, \ldots, r)\,,$$

so ist $\Phi \in C^1[a,b]$, aber $\deg \Phi\big|_{[x_{i-1},x_i]} \leq 3$.

4.1 Spline–Interpolation

Bemerkung 4.2 (Kubische Spline–Interpolation)

Kubische Splines erreichen ähnlich wie zusammengesetzte Hermite–Interpolierende eine hohe globale Glattheit, jedoch mit deutlich niedrigerem Polynomgrad:

$$s \in C^2[a,b]\,, \quad s\big|_{[x_i,x_{i+1}]} \in \Pi_3\,.$$

Splines der *Ordnung* k: $s \in C^{k-2}[a,b]$, $s\big|_{[x_i,x_{i+1}]} \in \Pi_{k-1}$.